Introduction to Data Packages

PRESENTED BY
Rufus Pollock
HDX Context
West Africa: Ebola Outbreak

Cumulative Cases of Ebola: 27,987
Cumulative Deaths from Ebola: 11,299
Response Plan Coverage: 68.7%

People Receiving Food Assistance
Open Ebola Treatment Centers
Currently Affected Countries

WHO
Data - Explore - Aug 17, 2015

WHO
Data - Explore - Aug 17, 2015

OCHA FTS
Data - Explore - Jul 21, 2015
Hamburger Analogy

Presentation

CKAN

ETL
[Tabular] Data Packages
<table>
<thead>
<tr>
<th>year</th>
<th>admin1</th>
<th>admin2</th>
<th>admin3</th>
<th>admin4</th>
<th>admin5</th>
<th>admin6</th>
<th>func1</th>
<th>func2</th>
<th>econ1</th>
<th>econ2</th>
<th>fin_source</th>
<th>exp_type</th>
<th>transfer</th>
<th>approved</th>
<th>adjusted</th>
<th>executed</th>
</tr>
</thead>
<tbody>
<tr>
<td>2009</td>
<td>Central</td>
<td>01</td>
<td>01</td>
<td>Parliament</td>
<td>010</td>
<td>Central apparatus</td>
<td>01</td>
<td>01</td>
<td>111 Remuneration</td>
<td>111.00 Remuneration of Base Personnel</td>
<td>Excluding</td>
<td>30269300</td>
<td>30269300</td>
<td>27469184</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td>Central</td>
<td>01</td>
<td>01</td>
<td>Parliament</td>
<td>010</td>
<td>Central apparatus</td>
<td>01</td>
<td>01</td>
<td>112 Mandatory state</td>
<td>112.00 Mandatory state of Base Personnel</td>
<td>Excluding</td>
<td>35640000</td>
<td>35640000</td>
<td>5401021</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td>Central</td>
<td>01</td>
<td>01</td>
<td>Parliament</td>
<td>010</td>
<td>Central apparatus</td>
<td>01</td>
<td>01</td>
<td>113 Payment for</td>
<td>113.03 Office supplies,</td>
<td>Base Non-Excluding</td>
<td>1057700</td>
<td>1262700</td>
<td>1115740</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td>Central</td>
<td>01</td>
<td>01</td>
<td>Parliament</td>
<td>010</td>
<td>Central apparatus</td>
<td>01</td>
<td>01</td>
<td>113 Payment for</td>
<td>113.03 Office supplies,</td>
<td>Special Non-Excluding</td>
<td>15000</td>
<td>15000</td>
<td>5282.72</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td>Central</td>
<td>01</td>
<td>01</td>
<td>Parliament</td>
<td>010</td>
<td>Central apparatus</td>
<td>01</td>
<td>01</td>
<td>113 Payment for</td>
<td>113.05 Books and</td>
<td>Base Non-Excluding</td>
<td>150800</td>
<td>120800</td>
<td>92372.12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td>Central</td>
<td>01</td>
<td>01</td>
<td>Parliament</td>
<td>010</td>
<td>Central apparatus</td>
<td>01</td>
<td>01</td>
<td>113 Payment for</td>
<td>113.10 Drugs and</td>
<td>Base Non-Excluding</td>
<td>4400</td>
<td>4400</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td>Central</td>
<td>01</td>
<td>01</td>
<td>Parliament</td>
<td>010</td>
<td>Central apparatus</td>
<td>01</td>
<td>01</td>
<td>113 Payment for</td>
<td>113.11 Telecommunication</td>
<td>Base Non-Excluding</td>
<td>1257600</td>
<td>952600</td>
<td>783778.13</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td>Central</td>
<td>01</td>
<td>01</td>
<td>Parliament</td>
<td>010</td>
<td>Central apparatus</td>
<td>01</td>
<td>01</td>
<td>113 Payment for</td>
<td>113.13 Transportation</td>
<td>Base Non-Excluding</td>
<td>130000</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td>Central</td>
<td>01</td>
<td>01</td>
<td>Parliament</td>
<td>010</td>
<td>Central apparatus</td>
<td>01</td>
<td>01</td>
<td>113 Payment for</td>
<td>113.17 Current</td>
<td>Base Non-Excluding</td>
<td>650000</td>
<td>35000</td>
<td>21889</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td>Central</td>
<td>01</td>
<td>01</td>
<td>Parliament</td>
<td>010</td>
<td>Central apparatus</td>
<td>01</td>
<td>01</td>
<td>113 Payment for</td>
<td>113.18 Current repair of</td>
<td>Base Non-Excluding</td>
<td>239000</td>
<td>163000</td>
<td>86835.64</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td>Central</td>
<td>01</td>
<td>01</td>
<td>Parliament</td>
<td>010</td>
<td>Central apparatus</td>
<td>01</td>
<td>01</td>
<td>113 Payment for</td>
<td>113.20 Use of government</td>
<td>Base Non-Excluding</td>
<td>35000</td>
<td>65000</td>
<td>57912</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td>Central</td>
<td>01</td>
<td>01</td>
<td>Parliament</td>
<td>010</td>
<td>Central apparatus</td>
<td>01</td>
<td>01</td>
<td>113 Payment for</td>
<td>113.21 Trainings</td>
<td>Base Non-Excluding</td>
<td>300000</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td>Central</td>
<td>01</td>
<td>01</td>
<td>Parliament</td>
<td>010</td>
<td>Central apparatus</td>
<td>01</td>
<td>01</td>
<td>113 Payment for</td>
<td>113.22 Printing services</td>
<td>Base Non-Excluding</td>
<td>189300</td>
<td>114300</td>
<td>114204</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td>Central</td>
<td>01</td>
<td>01</td>
<td>Parliament</td>
<td>010</td>
<td>Central apparatus</td>
<td>01</td>
<td>01</td>
<td>113 Payment for</td>
<td>113.23 Representation costs</td>
<td>Base Non-Excluding</td>
<td>1610000</td>
<td>810000</td>
<td>335058.53</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td>Central</td>
<td>01</td>
<td>01</td>
<td>Parliament</td>
<td>010</td>
<td>Central apparatus</td>
<td>01</td>
<td>01</td>
<td>113 Payment for</td>
<td>113.30 IT and PC works</td>
<td>Base Non-Excluding</td>
<td>120000</td>
<td>20000</td>
<td>14373.88</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td>Central</td>
<td>01</td>
<td>01</td>
<td>Parliament</td>
<td>010</td>
<td>Central apparatus</td>
<td>01</td>
<td>01</td>
<td>113 Payment for</td>
<td>113.36 Compensation for</td>
<td>Base Non-Excluding</td>
<td>430500</td>
<td>430500</td>
<td>193208</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td>Central</td>
<td>01</td>
<td>01</td>
<td>Parliament</td>
<td>010</td>
<td>Central apparatus</td>
<td>01</td>
<td>01</td>
<td>113 Payment for</td>
<td>113.39 Compensation for</td>
<td>Base Non-Excluding</td>
<td>672000</td>
<td>272000</td>
<td>236280</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td>Central</td>
<td>01</td>
<td>01</td>
<td>Parliament</td>
<td>010</td>
<td>Central apparatus</td>
<td>01</td>
<td>01</td>
<td>113 Payment for</td>
<td>113.45 Goods and services</td>
<td>Base Non-Excluding</td>
<td>50000</td>
<td>20000</td>
<td>15729.51</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td>Central</td>
<td>01</td>
<td>01</td>
<td>Parliament</td>
<td>010</td>
<td>Central apparatus</td>
<td>01</td>
<td>01</td>
<td>114 Duty travels</td>
<td>114.01 Travels within the</td>
<td>Base Non-Excluding</td>
<td>1000</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td>Central</td>
<td>01</td>
<td>01</td>
<td>Parliament</td>
<td>010</td>
<td>Central apparatus</td>
<td>01</td>
<td>01</td>
<td>114 Duty travels</td>
<td>114.02 Travels abroad</td>
<td>Base Non-Excluding</td>
<td>2124800</td>
<td>933200</td>
<td>428432.75</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td>Central</td>
<td>01</td>
<td>01</td>
<td>Parliament</td>
<td>010</td>
<td>Central apparatus</td>
<td>01</td>
<td>01</td>
<td>116 Employer’s</td>
<td>116.01 Employer’s share of</td>
<td>Base Personnel Excluding</td>
<td>846700</td>
<td>846700</td>
<td>81368.66</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td>Central</td>
<td>01</td>
<td>01</td>
<td>Parliament</td>
<td>010</td>
<td>Central apparatus</td>
<td>01</td>
<td>01</td>
<td>135 Transfers to</td>
<td>135.20 Other transfers made</td>
<td>Base Non-Excluding</td>
<td>6199300</td>
<td>5899300</td>
<td>5631770</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td>Central</td>
<td>01</td>
<td>01</td>
<td>Parliament</td>
<td>010</td>
<td>Central apparatus</td>
<td>01</td>
<td>01</td>
<td>242 Purchase of</td>
<td>242.06 Purchase of fixed</td>
<td>Base Capital Excluding</td>
<td>1810100</td>
<td>1140100</td>
<td>244655.59</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td>Central</td>
<td>01</td>
<td>01</td>
<td>Parliament</td>
<td>010</td>
<td>Central apparatus</td>
<td>01</td>
<td>01</td>
<td>243 Capital repair</td>
<td>243.04 Capital refurbishment</td>
<td>Base Capital Excluding</td>
<td>600000</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td>Central</td>
<td>01</td>
<td>01</td>
<td>Parliament</td>
<td>010</td>
<td>Central apparatus</td>
<td>01</td>
<td>01</td>
<td>243 Capital repair</td>
<td>243.06 Capital refurbishment</td>
<td>Base Capital Excluding</td>
<td>1000000</td>
<td>283000</td>
<td>28228</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td>Central</td>
<td>01</td>
<td>01</td>
<td>Parliament</td>
<td>1001</td>
<td>Directorate</td>
<td>218 Building</td>
<td>01</td>
<td>01</td>
<td>111 Remuneration</td>
<td>111.00 Remuneration of</td>
<td>Base Personnel Excluding</td>
<td>2486000</td>
<td>2486000</td>
<td>2693486.5</td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td>Central</td>
<td>01</td>
<td>01</td>
<td>Parliament</td>
<td>1001</td>
<td>Directorate</td>
<td>218 Building</td>
<td>01</td>
<td>01</td>
<td>111 Remuneration</td>
<td>111.00 Remuneration of</td>
<td>Special Personnel Excluding</td>
<td>10000</td>
<td>10000</td>
<td>10000</td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td>Central</td>
<td>01</td>
<td>01</td>
<td>Parliament</td>
<td>1001</td>
<td>Directorate</td>
<td>218 Building</td>
<td>01</td>
<td>01</td>
<td>112 Mandatory state</td>
<td>112.00 Mandatory state</td>
<td>Base Personnel Excluding</td>
<td>488600</td>
<td>521100</td>
<td>521021.16</td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td>Central</td>
<td>01</td>
<td>01</td>
<td>Parliament</td>
<td>1001</td>
<td>Directorate</td>
<td>218 Building</td>
<td>01</td>
<td>01</td>
<td>112 Mandatory state</td>
<td>112.00 Mandatory state</td>
<td>Special Personnel Excluding</td>
<td>2300</td>
<td>2300</td>
<td>2300</td>
<td></td>
</tr>
</tbody>
</table>
### Tabular Data Package

- `datapackage.json`
  - metadata
  - csv schema

### CSV

<table>
<thead>
<tr>
<th>year</th>
<th>subdept1</th>
<th>subdept2</th>
<th>subdept3</th>
<th>subdept4</th>
<th>func1</th>
<th>func2</th>
<th>score1</th>
<th>score2</th>
<th>ref</th>
<th>source</th>
<th>type</th>
<th>transfer</th>
<th>approved</th>
<th>adjusted</th>
<th>executed</th>
</tr>
</thead>
<tbody>
<tr>
<td>2005</td>
<td>Central</td>
<td>101</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>111</td>
<td>111</td>
<td>111</td>
<td>Personnel</td>
<td>111</td>
<td>Personnel</td>
<td>30269300</td>
<td>30289300</td>
<td>30289300</td>
</tr>
<tr>
<td>2006</td>
<td>Central</td>
<td>101</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>112</td>
<td>112</td>
<td>112</td>
<td>Movable state</td>
<td>112</td>
<td>Movable state</td>
<td>5546400</td>
<td>5564000</td>
<td>5401021</td>
</tr>
<tr>
<td>2007</td>
<td>Central</td>
<td>101</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>113</td>
<td>113</td>
<td>113</td>
<td>Office supplies</td>
<td>113</td>
<td>Office supplies</td>
<td>10560000</td>
<td>15000000</td>
<td>11570000</td>
</tr>
<tr>
<td>2008</td>
<td>Central</td>
<td>101</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>113</td>
<td>113</td>
<td>113</td>
<td>Office supplies</td>
<td>113</td>
<td>Office supplies</td>
<td>15000000</td>
<td>15000000</td>
<td>526272</td>
</tr>
<tr>
<td>2009</td>
<td>Central</td>
<td>101</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>113</td>
<td>113</td>
<td>113</td>
<td>Books and Tools</td>
<td>113</td>
<td>Books and Tools</td>
<td>15000000</td>
<td>12000000</td>
<td>92027212</td>
</tr>
<tr>
<td>2010</td>
<td>Central</td>
<td>101</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>113</td>
<td>113</td>
<td>113</td>
<td>Drugs and Medicine</td>
<td>113</td>
<td>Drugs and Medicine</td>
<td>44500000</td>
<td>44500000</td>
<td>880000</td>
</tr>
<tr>
<td>2011</td>
<td>Central</td>
<td>101</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>113</td>
<td>113</td>
<td>113</td>
<td>Telecommunication</td>
<td>113</td>
<td>Telecommunication</td>
<td>13000000</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Data Package

Create, edit, and validate Data Packages.

Select Profile

Data Package

DataPackage

DataPackage is a format and specification for convenient delivery, installation and management of data. This is the base profile for DataPackage.

Name *

my-data-package

A lower-case string with only alphanumeric characters along with '.', '_' or '-' characters.

Title

My Data Package

A human-readable title for this package.

Description

{  "name": "my-data-package",  "title": "My Data Package"}

http://datapackagist.okfnlabs.org/
HDX (3W) Uploader / Checker

Editor
Data Type Inferer
Validate Data

CKAN + DataStore

Store schema on resource

All done in JS + HTML so *very* fast to iterate
Good Tables

Good Tables is a free hosted service for validating tabular data. Find out more

Let's get started

Add a data table (required)

- URL
- File

Add your data source here.

Choose URL or File to add a data source by your preferred method.

Want to see an example? Click on one of the two options below followed by the green 'Validate' button.

- Example: Invalid Structure
- Example: Invalid Schema (also sets a test schema to validate against)

Add a data schema (optional)

Select to validate this data against a JSON Table Schema (What's that?).

- URL
- File

Add your schema source here.

Choose URL or File to add a schema by your preferred method. Alternatively, choose a preset schema.

File format

Encoding

Schema presets

http://goodtables.okfnlabs.org/
Appendix
Data Containerization
Cargo Transport ~
1955
Manual, Slow, Costly
(and Dangerous)
Containerization
Data is Shipping
Pre-Containerization
“Standards”

Data Containerization

Connect different tools with a lightweight common interchange format
Standards
- Standards (a few, simple ones)
- Tools (primarily for integration)
- Documentation
- Datasets

3 minor "innovations"
- In formats - data package, tabular data package (both based heavily on pre-existing approaches e.g. CommonJS, DublinCore)
- In process - packaging as a distinct step
- Of approach - stick closely to existing tooling - prioritize integration
Data Package “Standards”

**Data package**: minimal wrapping, agnostic about the data its "packaging", designed for extension. Very flexible: can be used as a transport for any kind of data BUT limits integration and tooling. [http://data.okfn.org/standards/data-package/](http://data.okfn.org/standards/data-package/)

Tools

http://data.okfn.org/tools/

- Publish and manage data packages
  - Online services and offline libraries to get, create, manipulate and share data packages
- Data package support in existing tooling
  - Spreadsheets - Excel, Google Docs ...
  - DaaS - BigQuery, Redshift, ...
  - End-user tools - R, Tableau, ...
  - Code: Python, Ruby, Java, ...
Key Principles

1. Narrow Focus
   One (small) part of the data chain. One specific feature (packaging) or type of data (simple tabular)

2. The Web
   Build for the web. Use formats that are web "native" (JSON) and work naturally with HTTP (e.g. CSV streaming, CSV plain text nature). Distributed not centralized.

3. Existing Tools
   Work as easily as possible with existing tools. E.g. everyone has tools to use CSV and it's supported by almost every language.

4. Minimal, Clean, Sufficient Structure
   Only most essential metadata (but allow for extensibility). Schema information. No blank lines at the top of the table, no footnotes inlined in rows at the bottom.

5. Simplicity
   Use most basic formats. Add the minimum, do the least needed. Reflected in all of the above.
“Data Package Standards”

- Data Package
- Data Package Tabular Profile (#tabular)
- Data Package Geo Profile (#geo)
- More ...

Community

- Docs, Outreach, Events, ...

Tooling and Integration

Publish and Use [profile specific] Data Packages

Core Tools
- Command Line Tool
- Web Tool
- JS / Python / Ruby libs
- Validation
- Registry

Integration
- R / Excel / ...
- DBs, DaaS (Bigquery etc)
- OpenRefine, Dat etc
- Data API
- ...